

Creating a Test Information Profile for a Two-Dimensional Latent Space

Terry A. Ackerman

University of Illinois

In some cognitive testing situations it is believed, despite reporting only a single score, that the test items differentiate levels of multiple traits. In such situations, the reported score may represent quite disparate composites of these multiple traits. Thus, when attempting to interpret a single score from a set of multidimensional items, several concerns naturally arise. First, it is important to know what composite of traits is being measured at all levels of the reported score scale. Second, it is also necessary to discern that all examinees, no matter where they lie in the latent

trait space, are being measured on the same composite of traits. Thus, the role of multidimensionality in the interpretation or meaning given to various score levels must be examined. This paper presents a method for computing multidimensional information and provides examples of how different aspects of test information can be displayed graphically to form a profile of a test in a two-dimensional latent space. *Index terms: information, item response theory, multidimensional item response theory, test information.*

The purpose of most standardized achievement tests is to distinguish trait levels of examinees and thereby order individuals on a trait. Ordering examinees accurately requires that all items in a test discriminate between levels of the trait being measured. Problems arise when a test contains items that also discriminate between levels of traits other than the primary trait being measured. Unfortunately, because ordering is a unidimensional concept, examinees cannot be ordered on two or more traits at the same time, unless the ordering is based on, for example, a weighted sum of each trait being measured.

Most standardized achievement tests are hypothesized to be multidimensional (Harrison, 1986; Traub, 1983) and require the use of several abilities that can vary from item to item. Collectively, if items are capable of measuring different composites of abilities, the composites that are being measured should be assessed and this knowledge should be used to interpret not only the meaning of the observed score scale but also the consistency of this interpretation throughout the observable score range.

The first objective of this paper is to explain how to compute the information function for the two-dimensional item response theory (IRT) compensatory model. Then the information for a unidimensional composite that is a linear combination of two latent traits is evaluated. The second objective is to illustrate several graphical analyses that enable a psychometrician to formulate an information profile for a test. This profile helps to determine the composite of traits that is being measured, and how accurately those combinations of traits are being measured.

Multidimensional IRT

The Multidimensional Compensatory Model

For the purposes of this paper, multidimensional item response data will be assumed to be accurately modeled by a compensatory model given as

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 3, September 1994, pp. 257-275
© Copyright 1994 Applied Psychological Measurement Inc.
0146-6216/94/030257-19\$2.20

257

$$P(x_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{\exp[1.7(\mathbf{a}'_i \boldsymbol{\theta}_j + d_i)]}{1 + \exp[1.7(\mathbf{a}'_i \boldsymbol{\theta}_j + d_i)]} \quad (1)$$

where

- x_{ij} is the 0–1 score on item i by person j ,
- \mathbf{a}_i is the vector of item discrimination parameters,
- d_i is a scalar parameter that is related to the difficulty of the item, and
- $\boldsymbol{\theta}_j$ is the vector of trait level parameters for person j .

This model is considered to be compensatory because high θ levels on one dimension can partially compensate for low θ levels on the second dimension. Total compensation can only occur when all of the discrimination parameters for an item are equal in value. There is a discrimination parameter for each dimension, but only one difficulty parameter for an item. The difficulty parameter is indeterminate for each separate dimension (Reckase, 1985). This model does not impose a correlation between θ_1 and θ_2 . For graphical purposes, the θ_1 axis is always represented as orthogonal to the θ_2 axis because it is easier to work within a Cartesian coordinate system. However, in most cases in which the dimensions represent psychological traits, a scatterplot of estimated θ_1 and θ_2 for a group of examinees would indicate some level of correlation.

Information in Two-Dimensional IRT

Reckase & McKinley (1991) presented a method for computing information for a multidimensional IRT (MIRT) model. Green (1990) presented a revised version that takes into account the lack of local independence when the information is estimated for a particular direction. The method discussed here is an extension of Green's work.

The information function $I(\boldsymbol{\theta})$, for a multivariate sample X with individual parameters $\boldsymbol{\theta}$ can be written as

$$I(\boldsymbol{\theta}) = -E \left[\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2} \right], \quad (2)$$

where E denotes the expected value and L is the likelihood function of $\boldsymbol{\theta}$ (Kendall & Stuart, 1973).

The two-dimensional case of Equation 2 can be expressed as

$$I_{i,i'}(\boldsymbol{\theta}) = -E \left[\frac{\partial^2}{\partial \theta_i \partial \theta_{i'}} \ln L \right] \quad \forall i = 1, 2; \quad i' = 1, 2, \quad (3)$$

or in matrix form as

$$I(\boldsymbol{\theta}) = -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \theta_1^2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial \ln L}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln L}{\partial \theta_2^2} \end{bmatrix}, \quad (4)$$

where the likelihood function for Equation 1 can be written as

$$L(\mathbf{u} | \boldsymbol{\theta}) = \prod_{i=1}^n L(u_i | \theta_1, \theta_2) = \prod_{i=1}^n P_i^{u_i} Q_i^{1-u_i}, \quad (5)$$

where

- P_i is the probability of a correct response as defined in Equation 1,
- Q_i is the probability of an incorrect response or $1 - P_i$, and

u_i is the dichotomous (0,1) item score.
 The natural logarithm of Equation 5 then is given as

$$\ln L = \sum_{i=1}^n [u_i \ln P_i + (1 - u_i) \ln(1 - P_i)]. \quad (6)$$

The multidimensional information in Equation 4 can be written as

$$I(\theta) = \begin{bmatrix} \sum_{i=1}^n 2.89 P_i Q_i a_{1i}^2 & \sum_{i=1}^n 2.89 P_i Q_i a_{1i} a_{2i} \\ \sum_{i=1}^n 2.89 P_i Q_i a_{1i} a_{2i} & \sum_{i=1}^n 2.89 P_i Q_i a_{2i}^2 \end{bmatrix}. \quad (7)$$

(A complete derivation is provided in the Appendix.) For a single item or for n items that measure in exactly the same direction (i.e., $a_{1i} = a_{2i}$, $\forall i = 1, \dots, n$), Equation 7 is singular. This result, which can be generalized to higher ordered dimensions (i.e., dimensions beyond the first dimension), should be expected because in both cases the test would be, by definition, unidimensional. Hence, for $I(\theta)$ to be positive definite, there must be at least two items that measure different composites of two traits. It appears that a necessary, but not sufficient, condition for unidimensionality within an IRT setting is for the information matrix of the maximum likelihood estimate (MLE) to be singular.

Information for a Linear Composite of Two Underlying Traits

Assume that a single score is reported for a test that is actually capable of discriminating between levels of two distinct and probably correlated traits. Implicitly, the unidimensional θ scale represents a composite of the two latent traits— θ_1 and θ_2 . In two-dimensional IRT, once a particular composite of traits is selected, local independence does not hold with respect to this composite.

Reckase & McKinley (1991) defined MINF, as a variant of multidimensional information, beginning with the expression

$$I_i(\theta) = \frac{\left(\frac{\delta P_i(\theta)}{\delta \theta} \right)}{P_i(\theta) Q_i(\theta)}, \quad (8)$$

and then computed the m -dimensional directional derivative

$$\nabla_{\alpha} P_i(\theta) = P_i(\theta) Q_i(\theta) \sum_{k=1}^m a_{ik} \cos \alpha_k \quad (9)$$

for the numerator of Equation 8. The final form of MINF, for item i , in a (θ_1, θ_2) composite direction α , is

$$I_{\alpha}(\theta) = \frac{\left[P_i(\theta) Q_i(\theta) \sum_{k=1}^m a_{ik} \cos \alpha_k \right]^2}{P_i(\theta) Q_i(\theta)}. \quad (10)$$

Although useful, MINF is not the multidimensional information shown in Equation 7. Instead, MINF is essentially a multidimensional critical ratio (Lord, 1980, p. 69). MINF is a measure of how effective test score x is at discriminating between a trait level (θ_1, θ_2) and a trait level “close by” (θ'_1, θ'_2) along a line through (θ_1, θ_2) at angle α . MINF, when computed for more than a single item, is not a replacement for the information

measure of Equation 7 because it fails to take into account the lack of local independence (i.e., Equation 9 fails to properly account for the covariance that results from specifying a particular direction). It does, however, have the advantage of providing an approximation to the measurement precision for a single item for a particular composite. [Note that Reckase & McKinley (1991) used a logistic model for their form of the compensatory model (i.e., Equation 1) without the 1.7 scaling factor. Thus, MINF lacks the effect of having this scaling factor present.]

Junker (1992) cautioned against estimating this composite from a unidimensional perspective if the test is not essentially unidimensional (Stout, 1987, 1990) [e.g., using LOGIST (Wingersky, Barton, & Lord, 1982) or BILOG (Mislevy & Bock, 1990) to obtain an estimate of θ , when the complete latent space is defined by m traits, $(\theta_1, \theta_2, \dots, \theta_m)$]. Primarily because of the lack of local independence, linear composites of the m underlying traits cannot be estimated consistently, and subsequently the unidimensional MLE, $\hat{\theta}$, does not converge to the composite θ_c where

$$\theta_c = \beta_1\theta_1 + \beta_2\theta_2 + \dots + \beta_m\theta_m \tag{11}$$

for any choice of composite weights $\beta_1, \beta_2, \dots, \beta_m$ [and in particular Wang's (1986) reference composite].

However, if the test is not essentially unidimensional and the θ s are estimated using a multidimensional model that accounts for the complete latent space and thus reinstates local independence [e.g., using calibration programs such as TESTFACT (Wilson, Wood, & Gibbons, 1984), MIRTE (Carlson, 1987), or NOHARM (Fraser & McDonald, 1988)] then the fact can be used that the MLE of

$$\theta = \beta_1\theta_1 + \beta_2\theta_2 + \dots + \beta_m\theta_m \tag{12}$$

is equal to $\beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2 + \dots + \beta_m\hat{\theta}_m$, where $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ are the MLEs of the trait parameters (Zehna, 1966). Note that in this case the MLE does converge to θ_c .

For the current discussion, assume that the unidimensional trait level θ is equal to the linear composite, θ_c , defined by $\theta_c = \beta_1\theta_1 + \beta_2\theta_2$, in which θ_c represents a linear composite in the angular direction of α and $\beta_1 = \cos\alpha$ and $\beta_2 = \sin\alpha$. To find the information function for this user-selected composite, the information matrix (Equation 7) must first be inverted to obtain the asymptotic covariance matrix for $(\hat{\theta}_1, \hat{\theta}_2)$ (Lehmann, 1983, Theorem 4.1, pp. 429–430).

The inverse of Equation 7 can be written as:

$$\text{COV} = \begin{bmatrix} \frac{\sum PQa_2^2}{|I(\theta_1, \theta_2)|} & -\frac{\sum PQa_1a_2}{|I(\theta_1, \theta_2)|} \\ -\frac{\sum PQa_1a_2}{|I(\theta_1, \theta_2)|} & \frac{\sum PQa_1^2}{|I(\theta_1, \theta_2)|} \end{bmatrix}, \tag{13}$$

where $|I(\theta_1, \theta_2)|$ is the determinant of the information matrix given as

$$\sum 2.89PQa_1^2 \sum 2.89PQa_2^2 - (\sum 2.89PQa_1a_2)^2. \tag{14}$$

To compute the information for a particular composite, Equation 13 should be premultiplied by the row vector $[\beta_1 \beta_2]$ and postmultiplied by the column vector

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{15}$$

to find

$$\text{Var}(\beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2) \equiv \text{Var}(\hat{\theta}). \quad (16)$$

The reciprocal of the resulting scalar quantity,

$$(\cos\alpha)^2 \text{Var}(\hat{\theta}_1|\theta_1, \theta_2) + (\sin\alpha)^2 \text{Var}(\hat{\theta}_2|\theta_1, \theta_2) + 2(\sin 2\alpha)\text{Cov}(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) \quad (17)$$

represents multidimensional information for the MLE $\hat{\theta}_c$ of the specified composite $\beta_1\theta_1 + \beta_2\theta_2$.

Method

MIRT Item Vector Representation

Insight about the multidimensional nature of items and examinee traits can be obtained using MIRT models. The work of Reckase & McKinley (1991) that formally defined MIRT item characteristics provided a framework for examining the interaction between multidimensional items and the multidimensional trait distributions associated with groups of examinees.

In a two-dimensional latent trait space (e.g., math and verbal abilities), the a_i vector designates the composite of θ_1 and θ_2 that is being best measured. If $a_1 = a_2$, both dimensions would be measured with the same degree of accuracy. However, if $a_1 = 0$ and $a_2 = 1.0$, discrimination would only occur along the θ_2 dimension. The number of different trait composites being assessed can be made readily apparent. If all of the items measure exactly the same (θ_1, θ_2) composite (i.e., the same "direction" in the two-dimensional latent trait plane), the test would be strictly unidimensional. The more varied the composites being assessed by different items, the larger the off-diagonal terms in Equation 13 become.

Following Reckase & McKinley (1991), an item that requires two traits for a correct response can be represented in the two-dimensional latent trait space as a vector. The length of the vector for item i is equal to the degree of multidimensional discrimination, MDISC, given as

$$\text{MDISC}_i = (a_{1i}^2 + a_{2i}^2)^{1/2}. \quad (18)$$

MDISC is analogous to the unidimensional IRT model's discrimination parameter. The measurement direction of the vector in degrees from the positive θ axis is

$$\alpha_i = \arccos\left(\frac{a_{1i}}{a_{1i}^2 + a_{2i}^2}\right)^{1/2}. \quad (19)$$

This reference angle represents the composite of the (θ_1, θ_2) trait space that item i best measures. It is important to note that unless an item measures along an axis (e.g., $a_2 = 0$) it measures both dimensions to some degree. The item vector originates at, and is graphed orthogonal to, the $p = .5$ equiprobability contour. In the compensatory model (Equation 1), these equiprobability contours are always parallel. The signed distance, D , from the origin to the $p = .5$ contour is computed as

$$D_i = \frac{-d_i}{\text{MDISC}}. \quad (20)$$

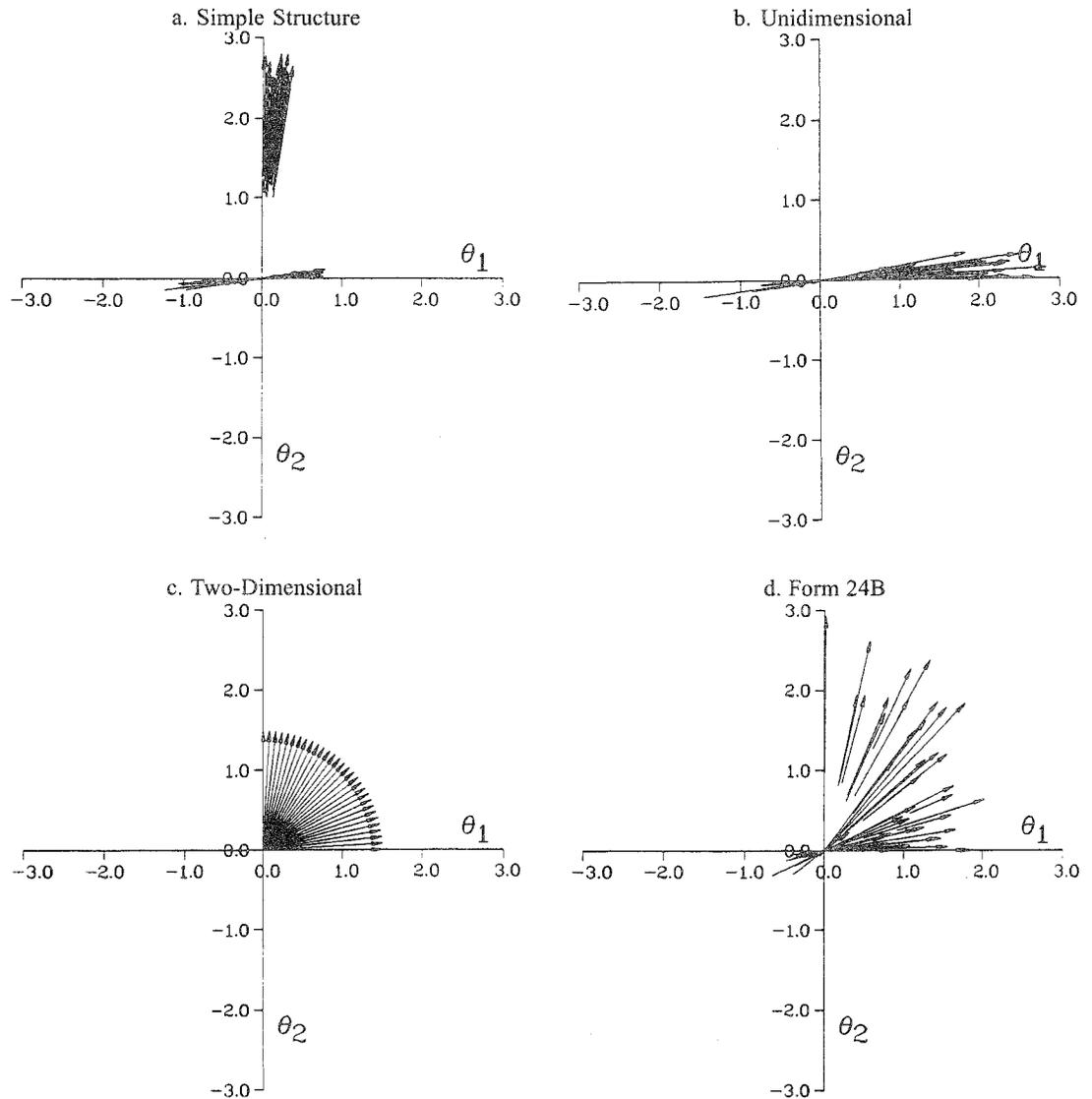
D_i is analogous to the unidimensional IRT difficulty parameter, b . Because the a s can never be negative, the item vectors can only lie in the third quadrant (representing easy items) or in the first quadrant (representing more difficult items); the first quadrant is bordered by the positive θ_1 and positive θ_2 axes and the remaining quadrants are numbered in a counterclockwise fashion.

Tests

To help illustrate the concept of multidimensional information, four different tests were examined. The

first three consisted of 30 items each. The two-dimensional item parameters generated for these tests were specifically contrived to portray three different tests: (1) a test with simple structure (i.e., half the items measured θ_1 and half measured θ_2), (2) a unidimensional test, and (3) a test in which the items measured different (θ_1, θ_2) composites from 0° to 90° in 3° increments (two-dimensional). The fourth test contained 40 items from the Math Usage Test of the ACT Assessment's Form 24B. Reckase (1985) reported that the parameters for Form 24B were two-dimensional estimates. Note that comparisons between tests assumed that the parameter estimates were all on the same scale. The item vector plot for each test is displayed in Figure 1.

Figure 1
Multidimensional Two-Parameter Logistic Item Vectors for Four Tests



The simple structure test (Figure 1a) might be created by item writers who, in an attempt to make items more difficult, require a different trait or an additional trait (e.g., placing an algebraic problem within the context of a story, thus requiring an additional skill of reading comprehension to solve the problem). The generated unidimensional test (Figure 1b) was composed of items whose vectors fell within a 10° sector of the θ_1 axis. The items for the two-dimensional test (Figure 1c) had equal MDISC and d values. The a_1 and a_2 parameters were varied to provide measurement for an equal “spread” of composites from 0° to 90° in 3° increments. The Form 24B (Figure 1d) item vectors demonstrated how varied the composite skills being measured by a standardized test can actually be: Items appear to be capable of distinguishing between trait levels for a full range of (θ_1, θ_2) composites.

Results

Test Response Surfaces

A contour plot of the test response surface (the two-dimensional analogue to the test response function in unidimensional IRT), computed from Equation 1, relates the “true score” scale to the latent trait plane (Figure 2). Each contour indicates the (θ_1, θ_2) combination that would have the expected observed score (“true” score) indicated by the number on the contour. Contour plots assist psychometricians in describing the meaning of the observed score scale in terms of the underlying latent traits.

Figure 2b shows that the observed score scale for the unidimensional test represents differences between levels of only θ_1 throughout the observed score range. Both the two-dimensional test (Figure 2c) and Form 24B (Figure 2d) seem to distinguish between levels of a composite skill denoted by an approximate equal weighting of θ_1 and θ_2 . The meaning of the observed score scale is not as consistent as in the unidimensional case because the contours are not strictly parallel, especially for extremely low and extremely high true scores. The worst case scenario, with respect to consistency of the interpretation of the true score scale, appears to be the simple structure test (Figure 2a). True score categories 1–12 (for examinees in the third quadrant) seem to reflect differences in primarily θ_1 , whereas score categories 16–28 (for examinees in the first quadrant) depict differences in mostly θ_2 .

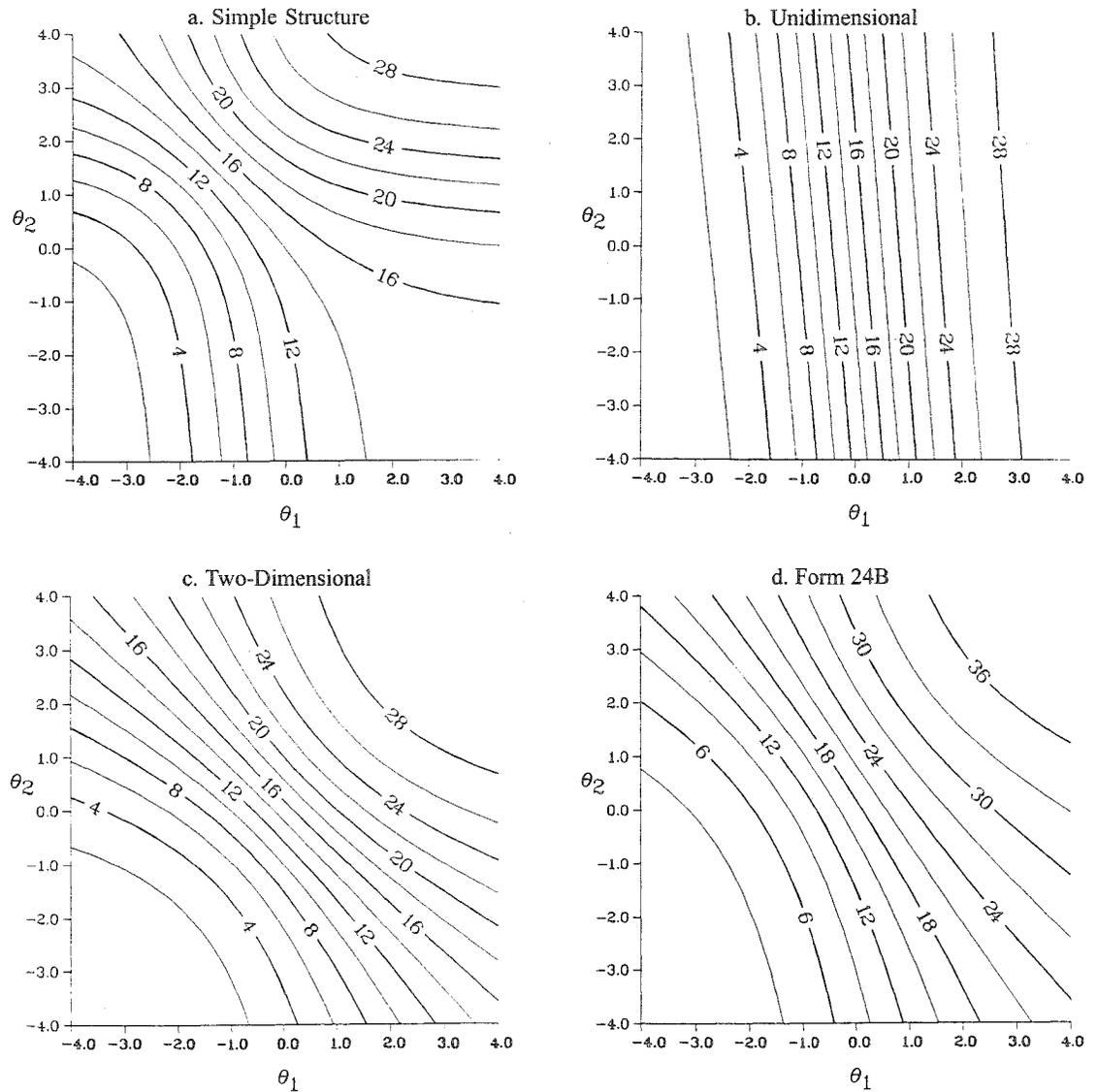
Test Information

The test information profile consists of a series of different graphs, each revealing how well a test is capable of measuring different composites of two-dimensional skills for examinees “located” throughout the latent trait space. The intent of such a graphical analysis is to clarify what a test measures. This knowledge then can be used to interpret different points along the observed score range.

Clamshell plots. A test information “clamshell” plot (Reckase & McKinley, 1991) shows how various composite skills are being measured throughout the two-dimensional trait plane. In the clamshell plot the test information value is represented as a vector. The longer the vector, the greater the measurement precision. Information is computed at each point in a 7×7 grid [for $(\theta_1, \theta_2) = -3.0$ to $+3.0$ in increments of 1 unit] in 10 different directions (for 0° to 90° in 10° increments). The individual clamshells identify which direction(s) provides the most information for various regions of the latent trait plane. A clamshell plot for each test is displayed in Figure 3.

The longest information vector at any given trait level in Figure 3 will be in the direction that is orthogonal to the nearest equitruer-score contour line shown in Figure 2. The scale of information is provided at the bottom of each plot. The unidimensional test (Figure 3b) provided the least amount of information, and Form 24B provided the greatest (Figure 3d). This result was expected because Form 24B had 10 more items than each of the other three tests. The simple structure plot (Figure 3a) shows the composite of skills that would be best measured in each quadrant. For example, in the first quadrant, examinees were measured primarily on θ_2 . In the second quadrant, it appears that the test measured all composites approximately

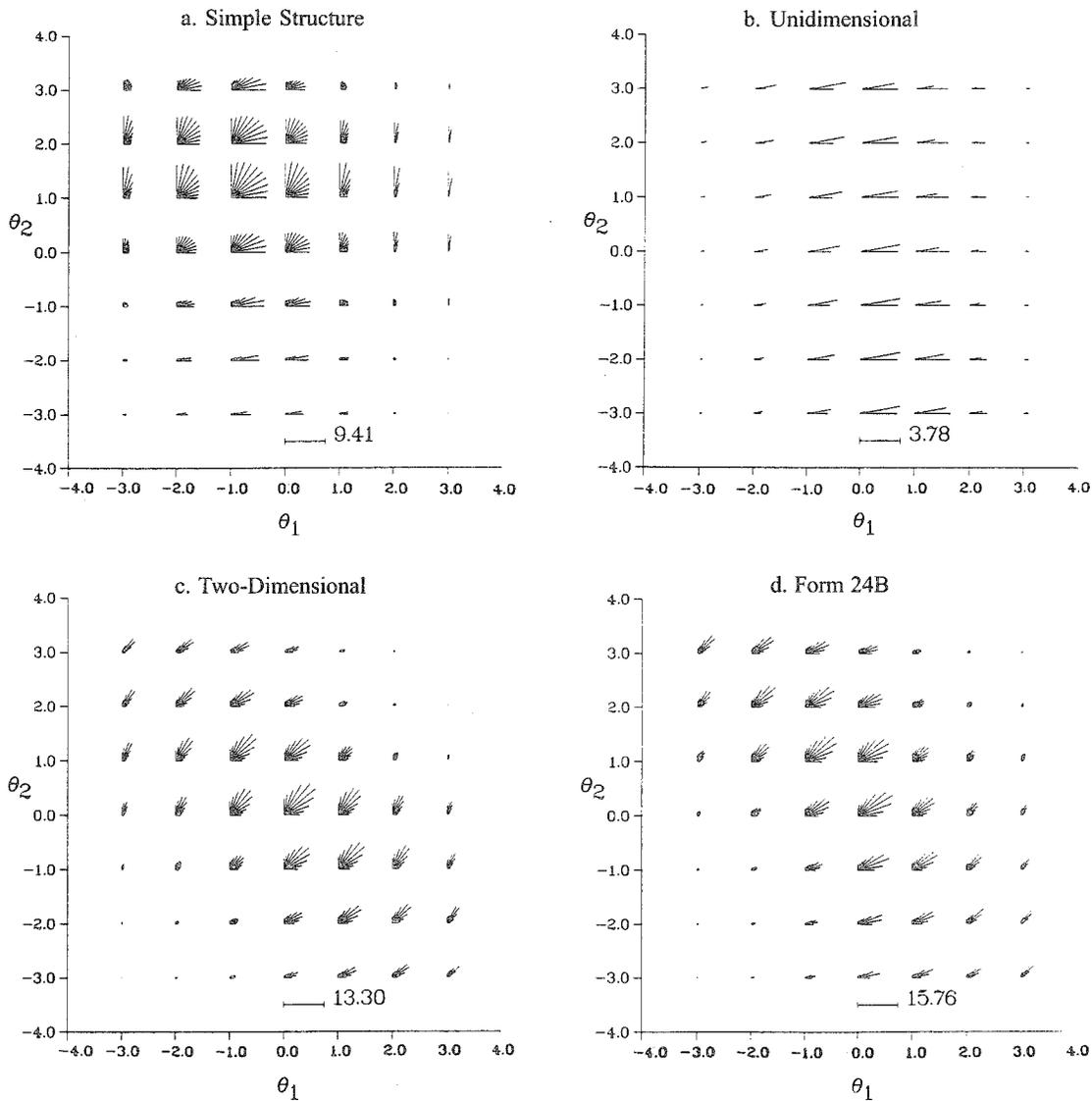
Figure 2
 Contour Plot of the Test Response Surface for Four Tests



equally well. However, examinees in the third quadrant were only distinguished by levels of θ_1 , and in the fourth quadrant there was very little measurement precision in any direction. The clamshells for the two-dimensional test (Figure 3c) and Form 24B are very similar to each other.

Information directional plots. After examining the clamshell plots, it is useful to ask what the direction of maximum information is at each point and how variable the directions are throughout the latent trait plane. Figure 4 shows the direction of maximum information for each of the 49-point grids shown in the clamshell plots. In this type of plot, the size of the print indicating the angle of the direction of maximum information is proportional to the amount of information (i.e., the larger the type size, the greater the

Figure 3
 Multidimensional Test Information Vectors at 49 Selected Trait Levels for Four Tests

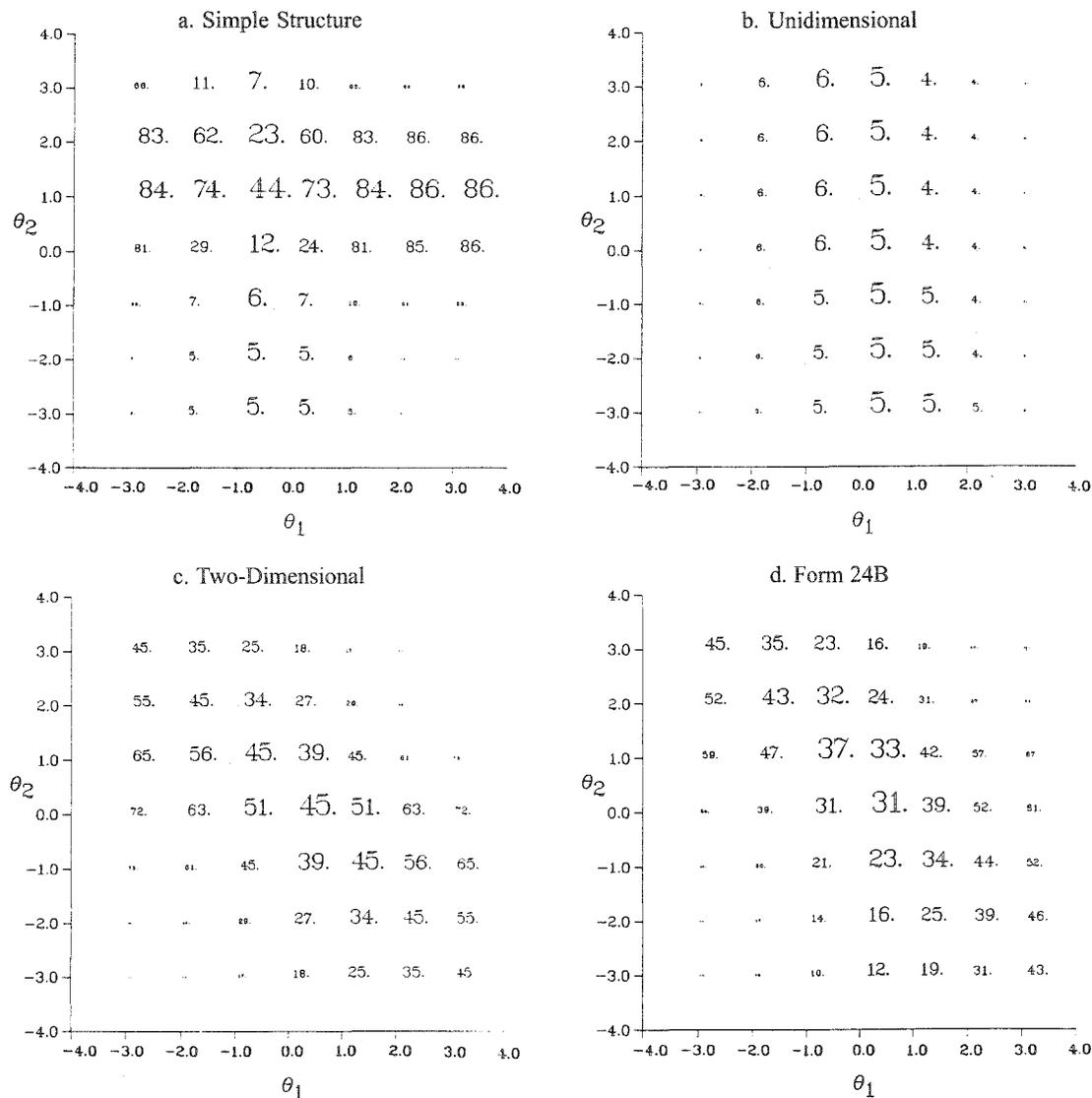


amount of information for the composite indicated by the angle specified by the number at that point).

If only a single score is reported for a test, it would be desirable to have each level of the resulting scale represent different levels of the same skill (or some composite of multiple skills). Such a case is the unidimensional case (Figure 4b). Even though the amount of information was not the same, the composite that was measured best tended to be between 4° and 6° for most examinees throughout the trait plane. The worst case was the simple structure test (Figure 4a). For this test, the composite that was measured best varied greatly from quadrant to quadrant (e.g., the first and third quadrants) and even within a quadrant (e.g., in the second quadrant). The two-dimensional test (Figure 4c) and Form 24B (Figure 4d) also seemed

Figure 4

Angles of Maximum Information at 49 Selected Trait Levels for Four Tests



to be quite variable, although only at the extreme trait levels was there a great deal of variability in the direction of best measurement. However, not many examinees would be expected in these trait regions.

Information surface plots. There is an information surface for each composite direction from 0° to 90°. The height of the surface, indicating the amount of information, at each of the 49 grid points for a 30° composite direction, for example, would be equal to the value indicated by the length of the vector in the 30° direction in the clamshell plot, at each trait level. For example, plots of the information surfaces for the simple structure test in the 0°, 30°, 60°, and 90° composite directions are plotted in Figure 5. The four surfaces can be compared to the length of the 0°, 30°, 60°, and 90° composite direction vectors shown in

Figure 5

Test Information Surface and Corresponding Contours for the 5° Composite Direction for the Unidimensional Test

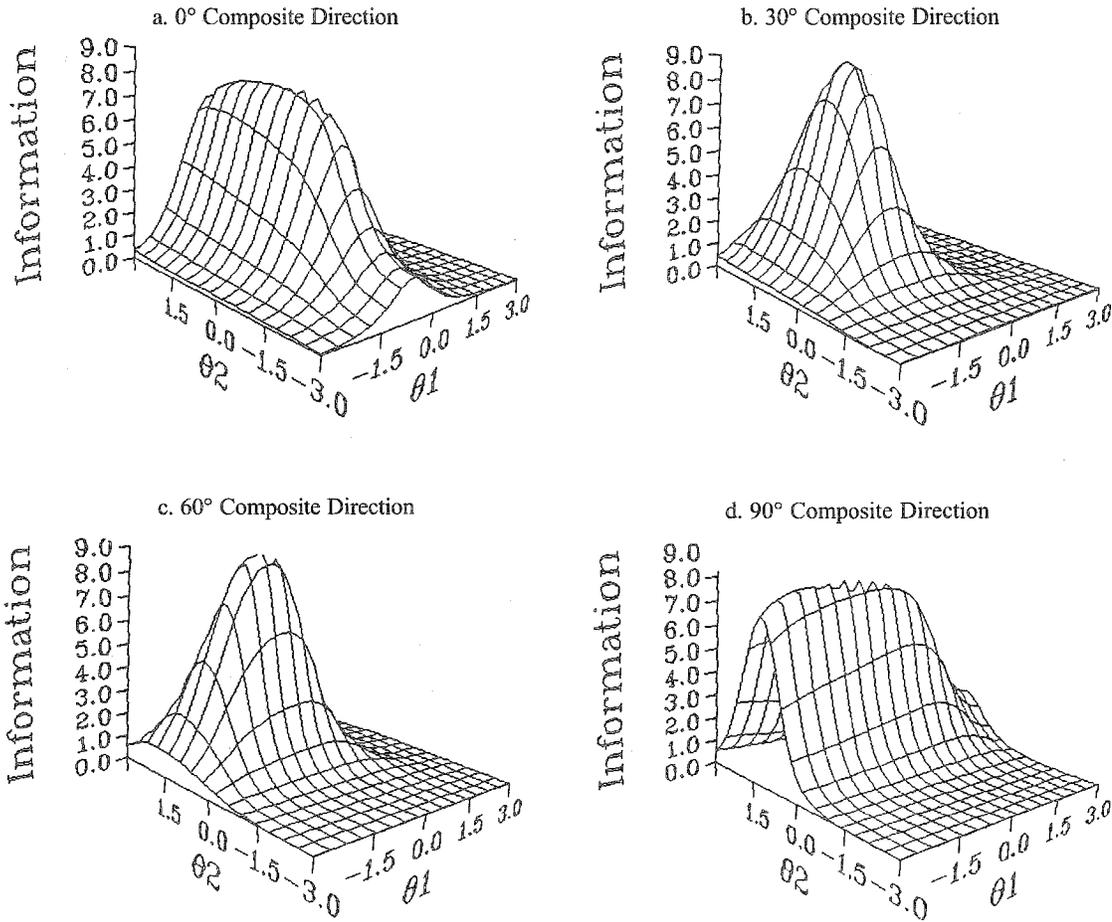


Figure 3a. Specifically, in the 0° direction the longest vectors were in a vertical band from $\theta_1 = -3.0$ to 0.0 and from $\theta_2 = -3.0$ to 3.0. This corresponds to the location of the ridged surface shown in Figure 5a. For the 30° and 60° directions, the clamshell plot indicates that most of the information was in the second quadrant. This is reflected in terms of the information surface plots in Figures 5b and 5c. Finally, note the ridge in the surface representing the 90° composite direction. This ridge corresponds to the band from $\theta_1 = -3.0$ to 3.0 and $\theta_2 = 0.0$ to 3.0 where the 90° vectors were the longest.

Information characteristic curves. At each of the 49 selected trait levels it is also important to know the rate at which the amount of information decreases when moving away from the direction of maximum information. Specifically, if the direction of maximum information is 50°, it would be interesting to know what the information would be at 55° or 45°. Whether the amount of information drops off quickly or gradually when moving away from the direction of maximum information is also of interest. To resolve these questions, a fourth plot can be examined.

The amount of information for composites from 0° to 90° in 1° increments was computed at each of the

49 selected trait levels. Plots displaying the 49 curves then were created from this information, and are shown in Figure 6. The vertical axis represents the amount of information, and the horizontal axis denotes the angular composite from 0° to 90° . By examining the variability of the height of each of the curves for any one angle, the degree of homogeneity of information for this composite can be assessed at the 49 different locations throughout the trait plane. If the same composite were being assessed in all regions of the trait plane, the curves would be homogeneous. In the perfect consistency case they would be coincident. This consistency was achieved only for the unidimensional test (Figure 6b). The collection of items on this test measured a composite in the direction of 5° – 6° from the θ_1 axis. The amount of information dropped off dramatically as the angle of the composite increased. This was not true for the other tests.

Figures 6a, 6c, and 6d show that for each of the other tests there was a range of curves with different maximums, some of which were in quite different directions. For example, the plot for the simple structure test (Figure 6a) shows that several curves peaked near 0° , whereas several other curves peaked between 80° and 90° . The degree of composite measurement consistency was quite low. This implies that depending on where examinees lie in the latent trait plane, the types (or composites) of skills on which they are being measured differ greatly.

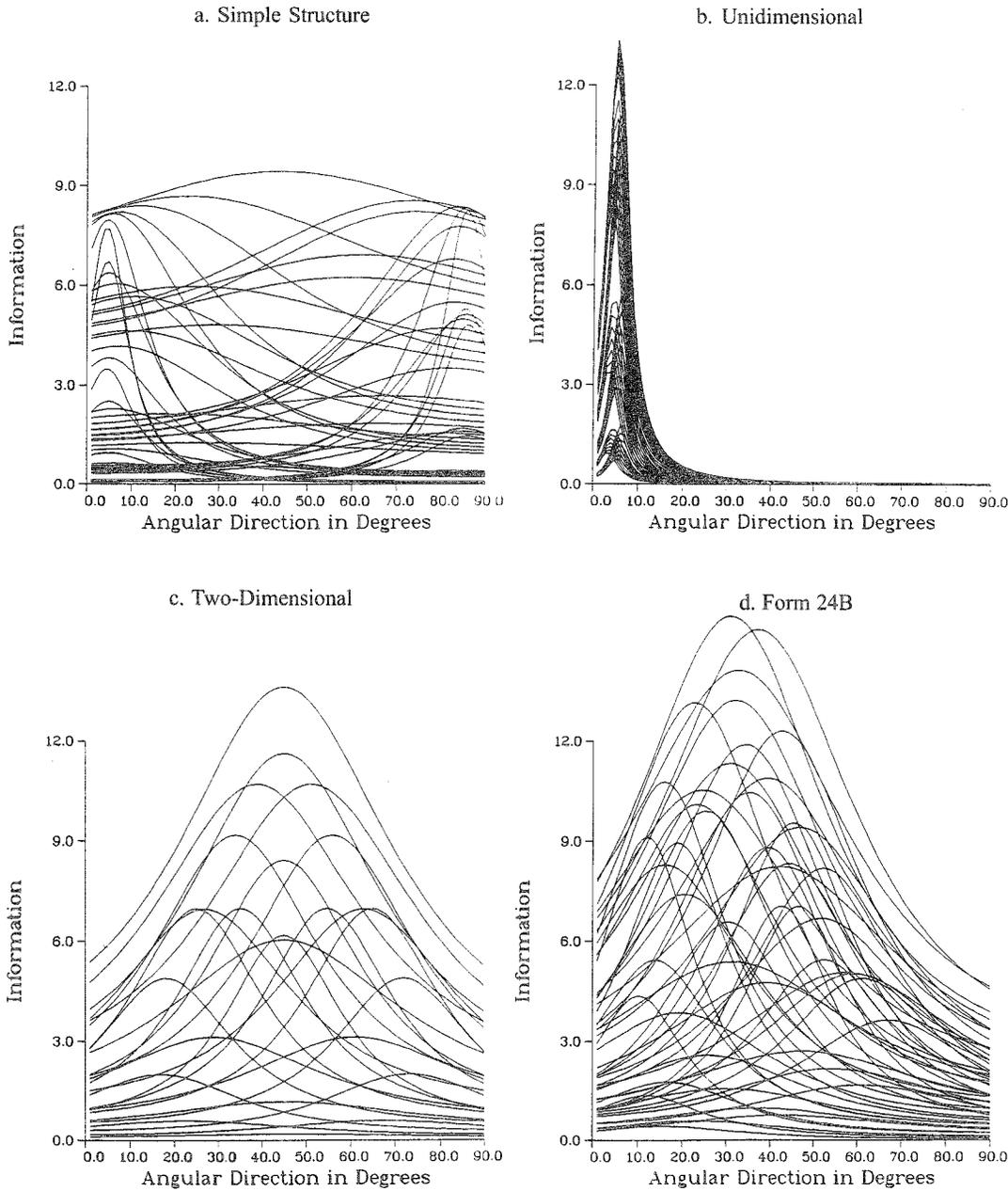
Directional composite correlations. Another plot indirectly reveals the relationship between measurement precision of different composites and the observed score (number-correct) scale. To create this plot, 2,000 examinees were generated randomly from a bivariate normal distribution ($\sigma_{\theta_1} = \sigma_{\theta_2} = 1.0$ and $\rho_{\theta_1, \theta_2} = .4$) for each of the four sets of item parameters (including Form 24B). The trait plane then was divided into octants and the correlation between examinees' observed scores and a linear composite of their (θ_1, θ_2) levels was computed for each composite representing 0° (only θ_1) to 90° (only θ_2) for examinees in each octant.

The direction corresponding to the linear composite that yielded the largest correlation was determined. This direction characterizes the composites of two-dimensional skills that relate best with the observed score for individuals in the area of the latent trait plane. A plot showing the composite direction with the highest correlation for each octant is shown for each test in Figure 7. The size of the font used to indicate each angle is directly proportional to the ratio of the squared correlation for the indicated angle with the maximum squared correlation achieved among the octants (i.e., the largest font is used for the angle having the greatest correlation between observed score and the indicated composite).

A moderate degree of similarity might be expected between each pair of corresponding plots in Figures 4 and 7. That is, the direction of maximum information indicated at the various grid points in Figure 4 should be similar to the direction that has the greatest correlation with the number-correct score. This did not appear to happen. In the first quadrant of the simple structure test (Figure 7a), the direction of maximum correlation was less than the composite direction of maximum information. The opposite was true in the third quadrant: The direction of maximum correlation was greater than the direction of maximum information. Surprisingly, Figure 4a indicates that there was relatively little information being provided in the third quadrant, yet Figure 7a implies a relatively large correlation between the observed scores and the latent trait composites in the indicated directions for examinees in this quadrant. The differences between Figures 4a and 7a might be due to the way the examinees were grouped in computing Figure 7a, which differs from a uniform spread of the 49 grid points in Figure 4. The range of indicated directions is 66° between examinees in the second and fifth octants.

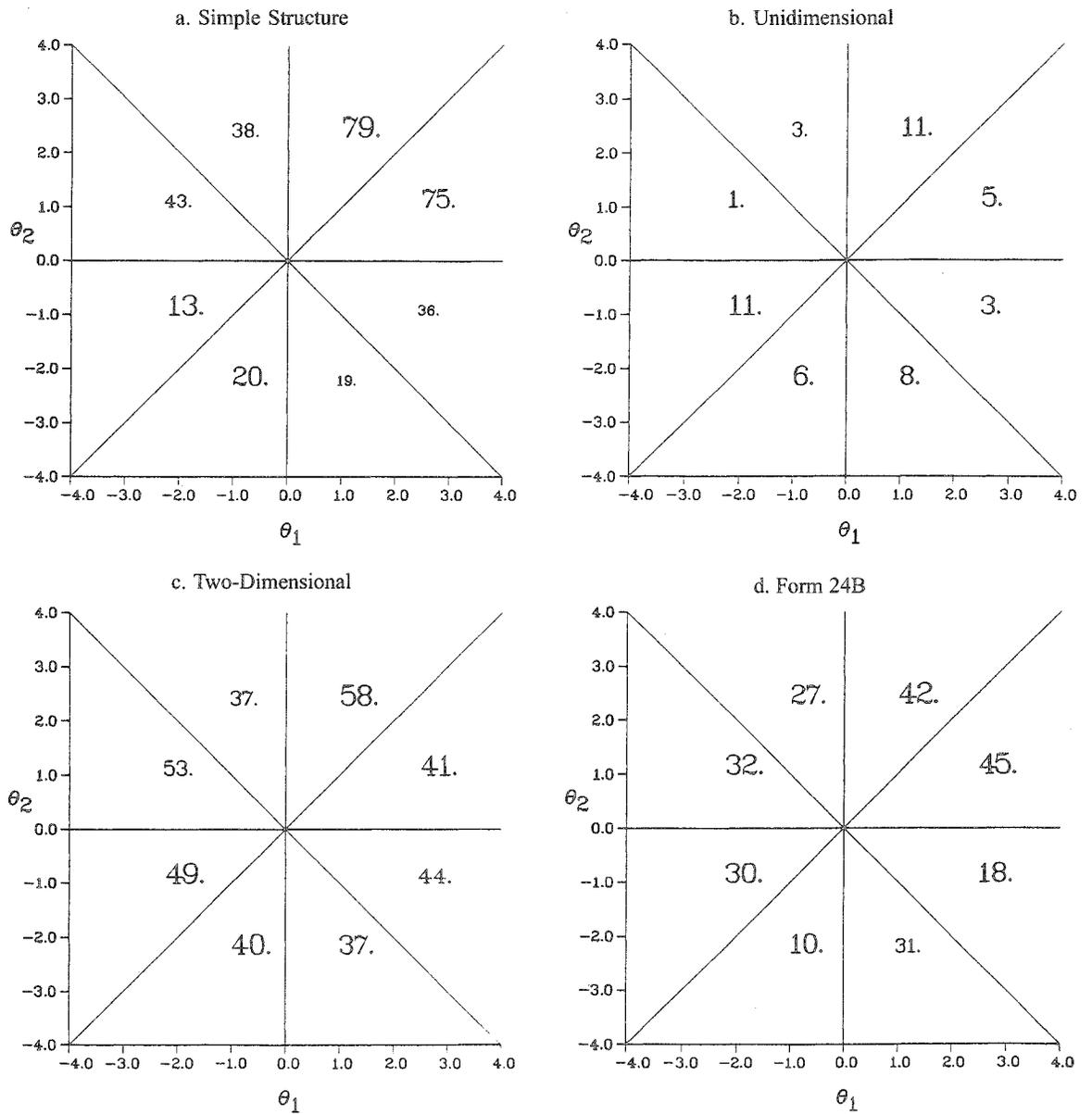
As in Figure 7a, Figures 7b, 7c, and 7d reveal inconsistency between the composite directions and the maximum correlation for the different octants. For the two-dimensional test, the symmetry that existed in Figure 4c does not exist in Figure 7c. The unidimensional test had the smallest range (10°) of composites having the maximum correlation. Figure 7d reveals that for Form 24B the greatest differences between examinees were in the second and sixth octants (42° and 10° , respectively).

Figure 6
Test Information Functions for 49 Selected Trait Levels for Different Composites
Ranging from 0° to 90° for Four Tests



True score composite directions. A final plot relates the true score value to the direction of maximum information. Throughout the observable score range it is important that the interpretation attached to

Figure 7
 Angles Representing the Directional Composite That Yielded the Highest Correlation,
 r_{θ_1, θ_2} , for Generated Examinees in Each Octant



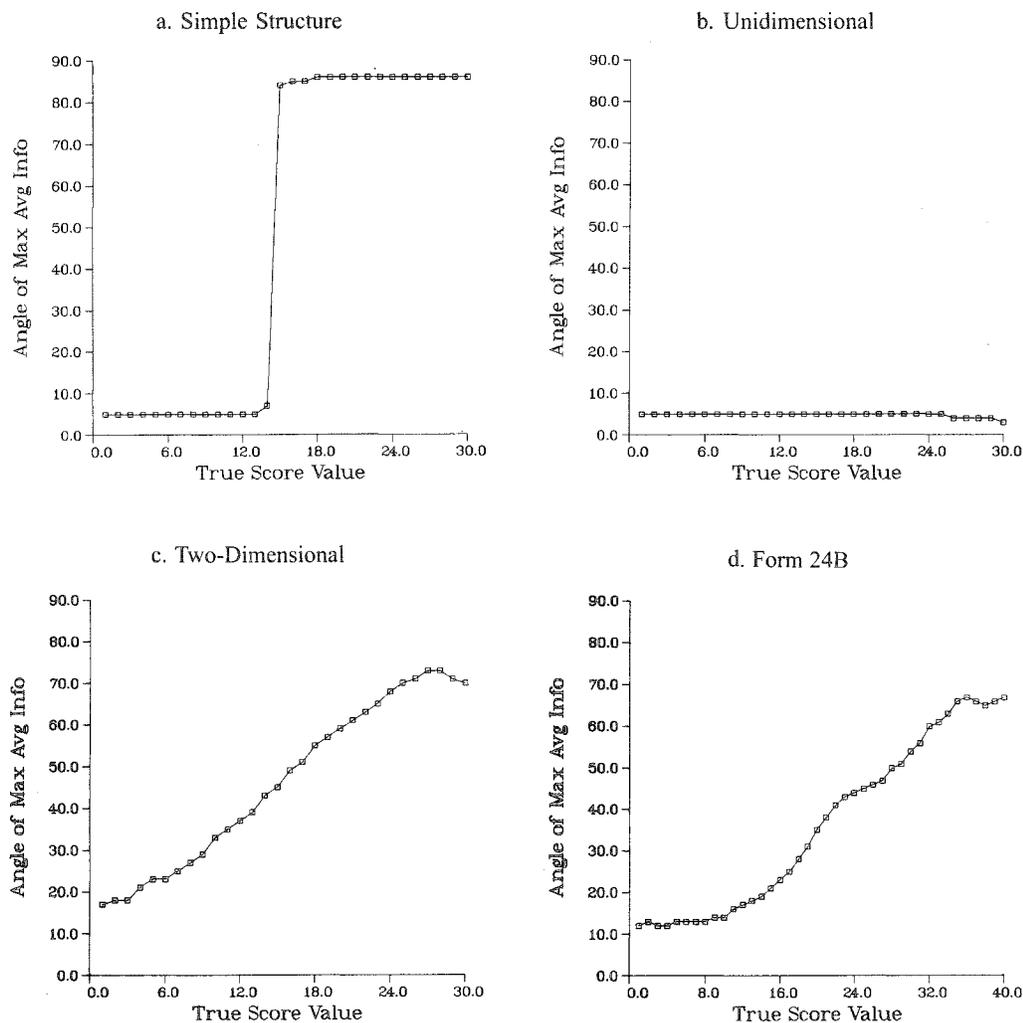
different score values be consistent. For example, in a 30-item test the difference between a score of 5 and a score of 6 should represent an increase in the same trait as the difference between a score of 25 and 26. To confirm this, several preliminary computations were performed. First, the "true" score, τ , for examinee j was computed for a 31×31 grid $[(\theta_1, \theta_2) = -3.0 \text{ to } 3.0 \text{ in } .2 \text{ increments}]$ of trait levels. τ is defined as

$$\tau_j = \sum_{i=1}^n p_{ij} \quad (21)$$

For those traits that had the same value of τ (rounded to the nearest integer), the information was computed for all composites from 0° to 90° and then averaged for each direction. The direction with the largest average then was determined for each τ value. Finally, a plot was created that displayed the direction of maximum average information for each τ value. If the score scale has a consistent interpretation, the direction of maximum average information should be the same throughout the τ range. The plots displaying these relationships for each of the four tests are shown in Figure 8.

Figure 8b shows that only for the unidimensional test was the angle of maximum average information the same for each τ value. As expected, the plot for the simple structure test (Figure 8a) revealed that $\tau =$

Figure 8
 The Direction of Maximum Average Information for Each Expected Score



1 to 14 were based only on θ_1 , and $\tau = 15$ through 30 represented different levels of θ_2 . The plots for the two-dimensional test (Figure 8c) and Form 24B (Figure 8d) both suggest a lack of score scale consistency (i.e., low scores represent trait composites quite different from high scores).

Discussion

The different plots were intended to graphically provide insight into what a two-dimensional test measures throughout the latent space and hence throughout the observed score scale. The clamshell plots in Figure 3 indicate how much information was provided about different composites at selected points in the latent trait plane. This plot also provides a sense of the degree to which the assumption of unidimensionality is violated. Psychometricians need to be concerned if a test distinguishes between levels of many different composite traits (e.g., the two-dimensional test or Form 24B) but only a single score is reported.

The difference between the composite traits that were measured in the example tests is shown in Figures 4a–4d. If only a single score were reported, the ideal situation would be to have the same composite being measured best for all examinees in the latent space (e.g., the unidimensional test, Figure 4b). This would indicate consistency of meaning or interpretation to the full range of the observed score scale. A worst case scenario was the simple structure test (Figure 4a) in which high trait level examinees in the first quadrant were separated according to their level of θ_2 , and examinees in the third quadrant were separated according to their level of θ_1 .

The plots in Figure 6 also provide a sense of which composites were being best measured. There was a striking difference between the unidimensional test (Figure 6b) and the simple structure test (Figure 6a). The angular direction of the modes of the 49 curves is the information provided in Figure 4. What is important to see from Figure 6, however, is the kurtosis of the curves. For example, for the unidimensional test the curves were quite leptokurtic. The composite being best measured throughout the plane can be interpreted as being within a narrow angular range.

The plots in Figures 7 and 8 relate the linear composites of the underlying traits to the observed and expected score scales. Two things need to be noted about the type of plot displayed in Figure 7. First, the composite angle that has the highest correlation with the number-correct scores should be similar for each of the octants. If only a single test score is being reported, the composite direction yielding the highest correlation should be the same for examinees in each octant. The unidimensional test (Figure 7b) came closest to this. Second, the size of the correlation indicates the degree of agreement between the ordering of examinees on the number-correct score and on the linear combination of (θ_1, θ_2) in each octant relative to the highest correlation among the octants. For example, for Form 24B (Figure 7d) the correlations were similar in magnitude, but the composite directions were different. The correlations for the simple structure test were variable and were so for rather disparate composites.

Figure 8 depicts graphically the consistency of the best average composite that was measured for each value of the true score scale. If tests are detected as being multidimensional, it should be confirmed that the composite that is being best measured is the same for all values of that score range.

One important question that needs to be studied further is how large a difference in degrees should be considered an important, substantive difference for interpreting the observed score scale. It might be interesting to divide the latent trait plane into concentric circles rather than octants and repeat the analysis.

Conclusions

The purpose of this paper was to demonstrate graphical ways to use multidimensional information to assess the composite of traits being measured by a single reported score when multiple traits are being measured (i.e., essential unidimensionality does not exist). It is important that those responsible for reporting test scores be able to interpret the observed score scale. If a test is multidimensional, but only a single

score is being reported, the composite that is best measured must be identified. For the sake of uniform interpretation it is also important that component traits be measured to the same degree of accuracy throughout the observable unidimensional score scale range.

The graphical approach presented here was predicated on the assumption that the test data fit the two-dimensional IRT compensatory model given in Equation 1. Before the plots to create a test information profile can be constructed, it must be determined that the response data are indeed two-dimensional and that estimating item parameters using Equation 1 is appropriate. Also, the accuracy of the information plots and subsequent interpretations are only as good as the item parameter estimates from which they were generated.

The information function computed at various trait levels in different composite directions describes the measurement precision of the maximum likelihood estimate of the latent traits. There is not a simple transformation from these underlying traits to an examinee's observed score. Thus, all except the last set of plots (Figure 8) require that conclusions be drawn about the true score and observed score based on evidence provided about the precision of maximum likelihood estimates of different linear composites of the underlying latent traits. Ideally, the test information profile in conjunction with a substantive cognitive analysis of the operations involved in solving the test items could help corroborate what a multidimensional test measures, especially if only a single score is being reported.

Appendix A

Extending the proof provided by Hambleton & Swaminathan (1985, p. 98) to the two-dimensional IRT case, the process of differentiating Equation 6 with respect to θ_1 and θ_2 can be approached by first applying the chain rule to yield

$$\frac{\partial \ln L}{\partial \theta_1} = \sum \frac{\partial \ln L}{\partial P} \frac{\partial P}{\partial \theta_1} \quad (22)$$

and

$$\frac{\partial \ln L}{\partial \theta_2} = \sum \frac{\partial \ln L}{\partial P} \frac{\partial P}{\partial \theta_2}. \quad (23)$$

The second derivatives then can be obtained by using the product rule:

$$\frac{\partial \ln L}{\partial \theta_1^2} = \sum \frac{\partial}{\partial \theta_1} \left(\frac{\partial \ln L}{\partial P_i} \right) \frac{\partial P_i}{\partial \theta_1} + \frac{\partial \ln L}{\partial P_i} \frac{\partial^2 P_i}{\partial \theta_1^2}, \quad (24)$$

$$\frac{\partial \ln L}{\partial \theta_1^2} = \sum \frac{\partial \ln L}{\partial^2 P_i} \left(\frac{\partial P_i}{\partial \theta_1} \right)^2 + \frac{\partial \ln L}{\partial P_i} \frac{\partial^2 P_i}{\partial \theta_1^2}, \quad (25)$$

and

$$\frac{\partial \ln L}{\partial \theta_2^2} = \sum \frac{\partial \ln L}{\partial^2 P_i} \left(\frac{\partial P_i}{\partial \theta_2} \right)^2 + \frac{\partial \ln L}{\partial P_i} \frac{\partial^2 P_i}{\partial \theta_2^2}. \quad (26)$$

Similarly, the derivative with respect to both θ_1 and θ_2 can be computed as

$$\frac{\partial \ln L}{\partial \theta_1 \theta_2} = \sum \frac{\partial \ln L}{\partial^2 P_i} \frac{\partial P_i}{\partial \theta_1} \frac{\partial P_i}{\partial \theta_2} + \frac{\partial \ln L}{\partial P_i} \frac{\partial P_i}{\partial \theta_1} \frac{\partial P_i}{\partial \theta_2}. \quad (27)$$

Next it is necessary to compute the first and second derivatives of the log of the likelihood with respect to P_i . Differentiating Equation 6 gives

$$\frac{\partial \ln L}{\partial P} = \frac{u}{P} - \frac{1-u}{1-P} \quad (28)$$

and

$$\frac{\partial^2 \ln L}{\partial P^2} = -\frac{u}{P^2} - \frac{1-u}{(1-P)^2}. \quad (29)$$

To complete the computation of $I(\theta_1, \theta_2)$, the expectations, using Equations 28 and 29, can be expressed as

$$E(u|\theta_1, \theta_2) = P_i, \quad (30)$$

$$E\left(\frac{\partial \ln L}{\partial P}\right) = E\left[\frac{u}{P_i} - \frac{1-u_i}{1-P_i}\right] = \frac{P_i}{P_i} - \frac{1-P_i}{1-P_i} = 0, \quad (31)$$

and

$$E\left(\frac{\partial^2 \ln L}{\partial P^2}\right) = E\left[-\frac{u}{P_i^2} - \frac{1-u_i}{(1-P_i)^2}\right] = -\frac{1}{P_i} - \frac{1}{1-P_i} = -\frac{1}{P_i Q_i}. \quad (32)$$

The components of Equation 4 now can be computed individually:

$$-E\left(\frac{\partial^2 \ln L}{\partial \theta_1^2}\right) = \sum \frac{\left(\frac{\partial P}{\partial \theta_1}\right)^2}{P_i Q_i}, \quad (33)$$

$$-E\left(\frac{\partial^2 \ln L}{\partial \theta_2^2}\right) = \sum \frac{\left(\frac{\partial P}{\partial \theta_2}\right)^2}{P_i Q_i}, \quad (34)$$

and

$$-E\left(\frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2}\right) = \sum \frac{\left(\frac{\partial P}{\partial \theta_1} \frac{\partial P}{\partial \theta_2}\right)}{P_i Q_i}. \quad (35)$$

When P_i is defined by the two-dimensional compensatory model (as in Equation 1), explicit expressions for the numerators of Equations 33–35 can be obtained as follows:

$$\frac{\partial^2 P_i}{\partial \theta_1^2} = 2.89 a_{1i}^2 P_i^2 Q_i^2, \quad (36)$$

$$\frac{\partial^2 P_i}{\partial \theta_2^2} = 2.89 a_{2i}^2 P_i^2 Q_i^2, \quad (37)$$

and

$$\frac{\partial P_i}{\partial \theta_1} \frac{\partial P_i}{\partial \theta_2} = 2.89 a_{1i} a_{2i} P_i^2 Q_i^2. \quad (38)$$

Finally, the multidimensional information in Equation 4 can be written as

$$I(\theta) = \begin{vmatrix} \sum_{i=1}^n 2.89 P_i Q_i a_{1i}^2 & \sum_{i=1}^n 2.89 P_i Q_i a_{1i} a_{2i} \\ \sum_{i=1}^n 2.89 P_i Q_i a_{1i} a_{2i} & \sum_{i=1}^n 2.89 P_i Q_i a_{2i}^2 \end{vmatrix}. \quad (39)$$

References

- Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (Research Rep. No. ONR87-2). Iowa City IA: American College Testing.
- Fraser, C., & McDonald, R. P. (1988). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: University of New England, Centre for Behavioral Studies.
- Green, B. F. (1990). *Notes on the item information function in the multidimensional compensatory IRT model* (Report No. 88-10). Baltimore MD: Johns Hopkins University, Psychometric Laboratory.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91-115.
- Junker, B. J. (1992, April). *Ability estimation in unidimensional models when more than one trait is present*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 2, 3rd ed.). New York: Hafner.
- Lehmann, E. L. (1983). *Theories of point estimation*. New York: Wiley.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
- Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.
- Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.
- Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver: Educational Research Institute of British Columbia.
- Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the ONR Contractors Conference, Gatlinburg TN.
- Wilson, D., Wood, R., & Gibbons, R. (1984). *TESTFACT: Test scoring, item statistics, and item factor analysis* [Computer program]. Mooresville IN: Scientific Software.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.
- Zehna, P. W. (1966). Invariance of maximum likelihood estimators. *Annals of Mathematical Statistics, 37*, 744.

Author's Address

Send requests for reprints or further information to Terry Ackerman, Department of Educational Psychology, University of Illinois, Champaign IL 61820-6990, U.S.A. Internet: ackerman@vmd.cso.uiuc.edu.